ANOMALY DETECTION IN SYSTEMS METRICS

VYTAUTO DIDŽIOJO UNIVERSITETAS
Informatikos fakultetas

AUTHORS:

Rimantė Kunickaitė
rimante.kunickaite1@vdu.lt

Dovilė Servaitė
dovile.servaite@vdu.lt

Gabrielė Jenciūtė
gabriele.jenciute@vdu.lt

Andrius Bumblauskas
andrius.bumblauskas@bluebridge.lt

Miglė Bučelytė
migle.bucelyte@bluebridge.lt

Aldas Glemža
aldas.glemza@bluebridge.lt

Tomas Krilavičius
tomas.krilavicius@vdu.lt

BlueBridge

CARD
CENTRE FOR APPLIED RESEARCH AND DEVELOPMENT

GOAL:

To detect anomalies in system metrics using time series methods.

DATASET:

In this research, metrics that represent the raw measurements of resource usage or behaviour that can be observed and collected throughout IT systems are analyzed.

- Number of threads
- CPU utilization
- Free disk space on E:
- Free memory

ANOMALY DETECTION:

Steps to be followed to find anomalies in time-series data:

1. Check whether the data is stationary or not. If the data is not stationary convert the data to stationary.
2. Split the data into training and testing set.
3. Fit (train) a time series model to the preprocessed training data.
4. Calculate the Absolute Error for each and every observation in the training data.
5. Find the threshold for the errors in the data.
6. Predict the observations in the testing data and calculate the errors.
7. If the error exceeds threshold, flag the observation as an anomaly.

MODELS:

- ARIMA
- TBATS
- PROPHET
- LSTM Neural Network
- Isolation Forest

Thresholds:
Static and dynamic thresholds are calculated according to the formulas:

$$TH_{static1} = \mu + k \cdot \delta;$$

$$TH_{dynamic1,t} = \mu_{\omega} + k \cdot \delta_{\omega};$$

$$TH_{static2} = mdn + k \cdot mad;$$

$$TH_{dynamic2,t} = \mu_{\omega} + k \cdot \delta_{\omega};$$

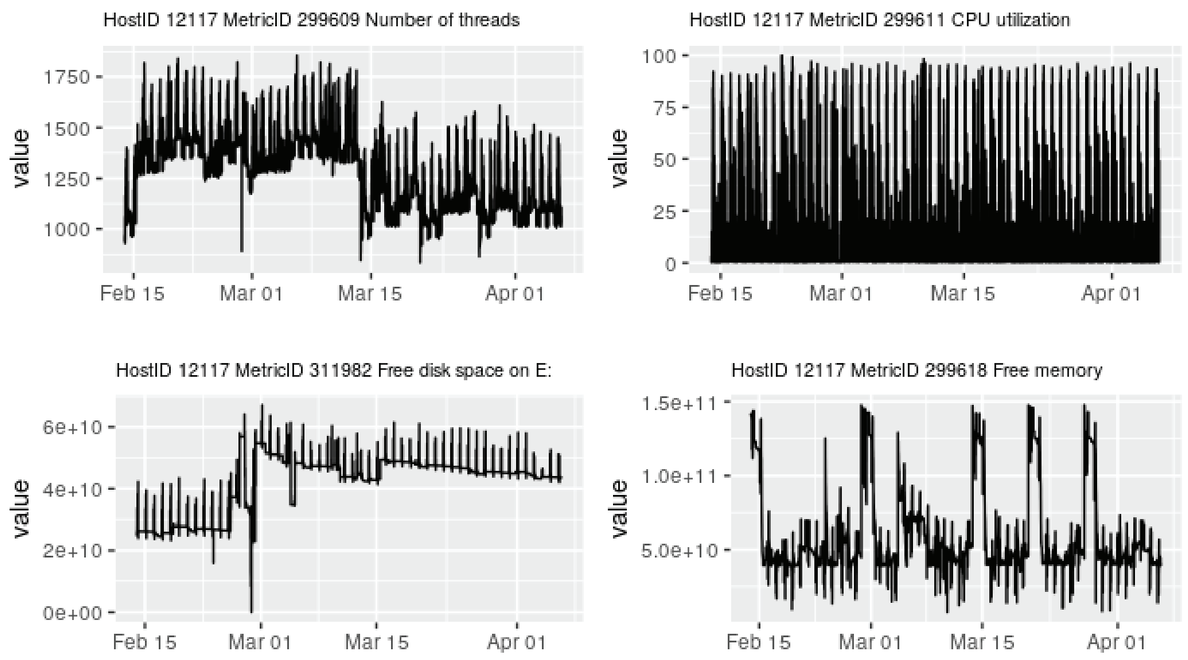
where μ - mean of prediction errors, δ - standard deviation of predictions errors, mad - median of predictions errors, ω - window size and k - coefficient of deviation.

RESULTS:

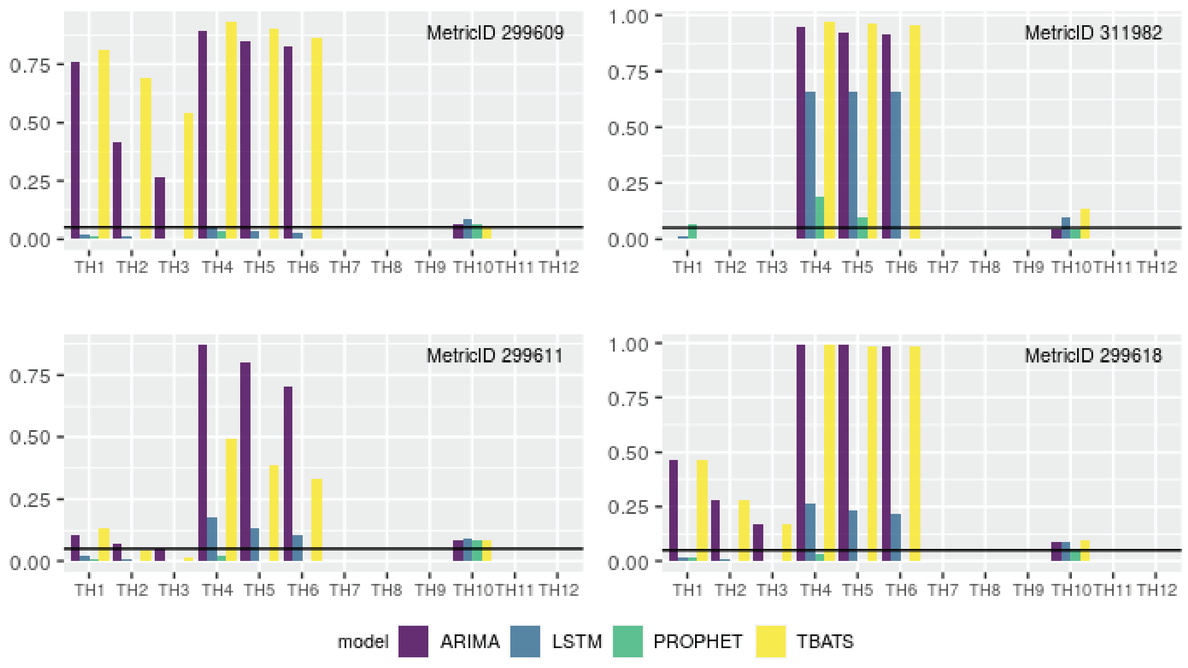
For each metric, the models were constructed separately using one week of data for training and one week of data for testing when the models were retrained on a daily basis. For the detection of anomalies, 12 threshold values were defined for each model:
TH1, TH2, TH3 $(TH_{static1,}$ when $k=3,5,7)$
TH4, TH5, TH6 $(TH_{static2,}$ when $k=3,5,7)$
and using a 5 minute sliding window TH7, TH8, TH9 $(TH_{dynamic1,t,}$ when $k=3,5,7)$

Table: Metrics descriptive statistics

NO.	METRIC ID	METRIC NAME	MIN	MEAN	MAX
1.	299609	Number of threads	835	1267,9493	1856
2.	299611	CPU utilization	0,1426	11,0932	99,9572
3.	311982	Free disk space on E:	2596864	41639118109	67171205120
4.	299618	Free memory	791746969	60481012936	147694555136

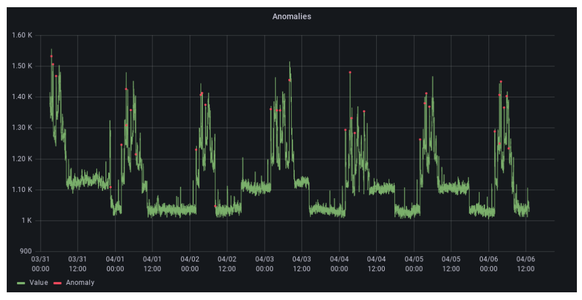


Picture: Percentage of detected anomalies by the models



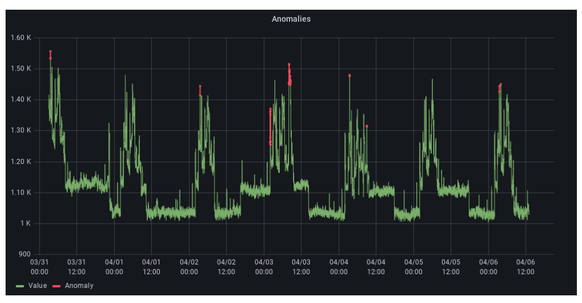
Picture: LSTM model detected anomalies in metric 299609 with threshold TH3

Picture: Number of detected anomalies by model in metric 280685 on 01/04/2021.



Picture: PROPHET model detected anomalies in metric 299609 with threshold TH5

Picture: Number of detected anomalies by static thresholds with LSTM model in metric 280685 on 01/04/2021.



FUTURE WORKS:

We plan to use multivariate time series models and contextual anomaly detection models to detect anomalies in IT systems metrics.